Methodologie d'analyse des resultats du modele de planification du reseau de transport a tres haute tension d'EDF

T. Cembrzynski

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 651

MÉTHODOLOGIE D'ANALYSE DES RÉSULTATS DU MODÈLE DE PLANIFICATION DU RÉSEAU DE TRANSPORT À TRÈS HAUTE TENSION D'EDF

Thierry CEMBRZYNSKI

Mars 1987

Méthodologie d'analyse des résultats du modèle de planification du réseau de transport à très haute tension d'EDF.

Analysis methodology for studying the results of EDF very high voltage grid planning model.

Thierry CEMBRZYNSKI
INRIA
Domaine de Voluceau
BP 105 - Rocquencourt -
78153 Le Chesnay Cedex - France -

Résumé

Nous présentons dans ces quelques pages la méthodologie d'analyse retenue pour l'étude des résultats du modèle de planification MEXICO. Cette méthode comporte deux phases:

-1: l'analyse macroscopique avec la classification automatique

-2: l'analyse du réseau éléctrique avec le Modèle Linéaire.

Abstract

We submit in some pages the analysis methodology used for studying the results of the planning model MEXICO. This method involves two parts:

-1: A macroscopic analysis with cluster analysis

-2: The grid analysis with the Linear Model.

SOMMAIRE

PREAMBULE

La fonction du réseau électrique à très haute tension est de transporter l'énergie depuis les centres de production (quelques centaines de centrales de divers types) jusqu'aux clients (plusieurs millions d'abonnés alimentés par l'intermédiaire de réseaux de distribution de tension inférieure). Le réseau est donc un système complexe, composé de quelques centaines de lignes de transport et de postes de transformation et d'interconnexion.

Le planificateur a donc pour mission de concevoir plusieurs années à l'avance, la structure du réseau de telle sorte qu'il puisse être exploité, le moment venu, dans les meilleures conditions de coût et de fiabilité.

Pour éclairer cette tâche, il dispose d'outils de calcul dont le principal est le modèle MEXICO. Cet outil permet d'estimer le comportement du réseau envisagé par le planificateur, et en particulier son aptitude à desservir la demande. Mais ce comportement du réseau est très aléatoire car il dépend de la disponibilité des ouvrages de production et de transport, du niveau de la demande, etc ...

C'est pourquoi il est nécessaire de simuler le fonctionnement du réseau dans un très grand nombre de situations tirées au hasard (quelques milliers). Les résultats du modèle sont des estimations de performances du réseau, telles que l'énergie non desservie, globale où par point de livraison, les transits d'énergie dans les ouvrages, etc ... Ces estimations sont obtenues comme moyennes de résultats de chaque situation.

De ce fait l'outil ne donne qu'une image "moyennée" de l'exploitation du réseau. Ceci fait de MEXICO un outil de mesure très apprécié du planificateur, mais non réellement un outil d'aide à la planification.Autrement dit, il permet de calculer l'amélioration des performances qu'apporterait le renforcement étudié par le planificateur et donc de décider de ce renforcement; mais il ne donne qu'une information moyenne sur le comportement du réseau qui n'en permet pas d'en déceler facilement les faiblesses, et non plus de suggérer d'autres renforcements.

Ainsi pour "faire" de MEXICO un réel outil d'aide à la planification, il faudrait le compléter d'un système automatique d'aide à l'analyse des résultats du modèle capable d'établir un diagnostic de l'état du réseau, d'en montrer les faiblesses , puis de suggérer des solutions de renforcements au planificateur.C'est cet objectif que nous poursuivons à long terme; la présente publication a pour but de montrer les travaux réalisés à ce jour.

Elle recouvre deux parties :

- l'analyse macroscopique des résultats du modèle (coûts d'exploitation) abordée par la Classification
- l'étude des éléments (liaisons électriques et groupes thermiques) du réseau par le Modèle Linéaire.

A : LA CLASSIFICATION

Rappelons tout d'abord quelques généralités. Il existe plusieurs familles d'algorithmes de classification automatique. Nous nous limiterons ici qu'aux techniques utilisées dans notre application : les algorithmes ascendants ou agglomératifs qui procèdent à la construction des classes par agglomération sucessive des éléments deux à deux et qui fournissent une **hiérarchie** de partitions des objets; et les algorithmes conduisant directement à des partitions; en particulier la méthode des **Nuées Dynamiques** (DIDAY), particulièrement intéressante dans le cas de grands ensembles de données.

A.a : Pourquoi la classification dans notre application ?

Nous avons opté pour la classification dans **l'analyse macroscopique** des résultats du modèle MEXICO afin de discriminer les problèmes pouvant se poser à l'exploitation du système électrique. Après avoir étudié diverses variables globales résultantes du modèle, nous avons décidé de retenir pour cette approche macroscopique les variables de surcoûts d'exploitation dus au réseau: (**VCNU,VCCH,VCHY,VCFG** : surcoûts d'exploitation Nucléaire, Charbon, Hydraulique,Fioul).

Celles-ci sont calculées comme la différence du coût d'exploitation avec réseau moins le coût d'exploitation sans réseau, ce dernier correspondant au coût minimal obtenu par empilement des disponibilités (tirées au hasard pour chaque situation ou aléa) des groupes de production, jusqu'à satisfaction de la consommation fixée par hypothèse.

Le réseau sera alors d'autant plus performant que pour chaque situation étudiée le surcoût d'exploitation global (nucléaire+charbon+hydraulique+fioul+défaillance) sera petit.

La classification permet ainsi de définir la **nature** du problème de réseau qui caractèrise chaque situation, on peut alors répondre à des questions aussi diverses que : " l'aléa **j** est il caractèrisé par un problème conduisant à une mauvaise utilisation de la production nucléaire, nécessitant alors du charbon ou du fioul ? "; ou encore que : " l'aléa **j** est il caractèrisé par un problème conduisant à des coupures importantes (défaillance) ? ". Ainsi permet elle la recherche de situations caractèristiques d'un problème particulier et également la **définition du problème global** caractèrisant l'échantillon étudié.

A.b : Les problèmes liés aux méthodes de classification

Pour la classification ascendante hiérarchique, on se heurte principalement à deux problèmes; l'un essentiellement pratique et l'autre plus théorique.

Le problème pratique est que cette méthode est difficilement utilisable pour le traitement de vastes recueils de données, car elle impose le calcul, le stockage et la mise à jour de la demi-matrice des distances entre les objets à classer, matrice dont la taille devient colossale dès que l'on traite quelques milliers d'objets

comme c'est le cas dans notre application. Le problème théorique réside quant à lui dans le choix, par l'utilisateur, à la fin de classification, d'une partition en q classes parmi la hiérarchie des partitions possibles, sans pour autant qu'il soit sûr que la partition obtenue soit effectivement optimale (en prenant comme critère à minimiser la variance intra-classes).

Pour les techniques d'agrégation autour de centres mobiles, les problèmes sont de nature différente.

Les algorithmes convergent toujours vers des **optimas locaux**. Le problème de la recherche d'une partition en q classes (en prenant comme critère à minimiser la variance intra-classes) , n'a pas encore donné lieu à un algorithme vraiment satisfaisant. Les partitions dépendent fortement des premiers centres choisis et également de l'ordre des éléments à classer. Ces algorithmes sont très efficaces pour décrire rapidement de vastes recueils de données, cependant ils imposent à l'utilisateur de **choisir a priori le nombre de classes** qu'il désire, et ce choix est bien souvent fort délicat, particulièrement lorsqu'on ne connait pas le contenu de ses données. Il existe bien des méthodes dites à nombre de classes variables BOULES,ISODATA mais l'utilisateur doit alors définir d'autres paramètres destinés à "piloter" l'élaboration de la partition ce qui en rend l'utilisation fort délicate.

A.c : La stratégie de classification utilisée dans notre application

Dans la mesure où l'outil d'analyse développé est destiné à des spécialistes de la planification des réseaux qui ne sont pas, en revanche des spécialistes de l'analyse des données, il convenait d'élaborer une méthode de classification fournissant en un essai –et avec un temps de calcul raisonnable– une partition de bonne qualité dans laquelle il leurs serait facile de définir les problèmes rencontrés.

Dans le cadre de notre étude nous avons alors retenu les Nuées Dynamiques avec cependant une différence importante, qui réside dans l'utilisation d'une classification ascendante hiérarchique (Méthode de Ward) pour à la fois "détecter" automatiquement un nombre de classes statistiquement significatives et sélectionner de bons noyaux initiaux. Le principal avantage de cette méthode est que l'utilisateur n'est pas obligé d'imposer à priori un nombre de classes, c'est au cours de la classification que la procédure le détermine automatiquement.

La stratégie suivie dans notre programme de classification est la suivante :

Etape 1 phase d'apprentissage : on effectue les Nuées Dynamiques sur un grand nombre de classes (trente classes pour être précis, ce qui est largement suffisant dans le cadre de notre application); et puisque le choix des noyaux de départ conditionne fortement la partition finale, on sélectionne initialement ceux-ci sous contraintes (cf :A.f).

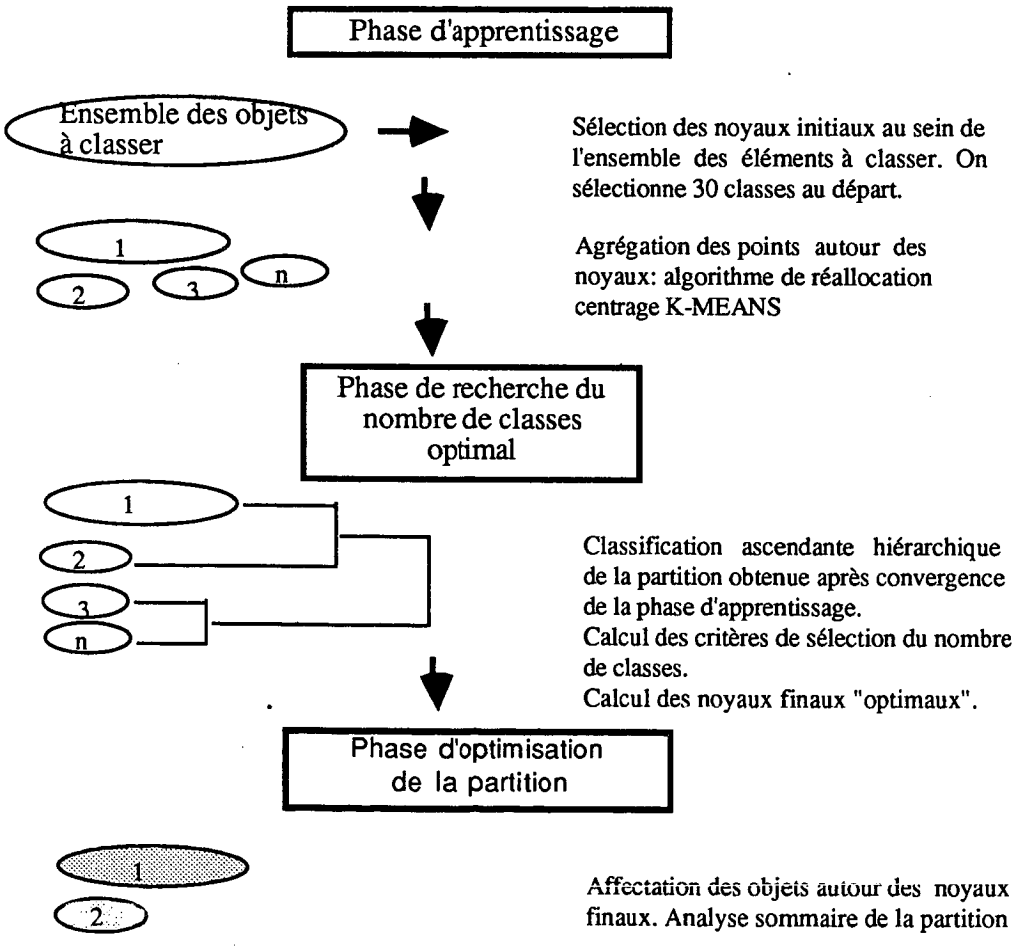
Etape 2 phase de recherche du nombre de classes : on effectue une classification ascendante hiérarchique (sur les centres des classes de la partition obtenue à la fin de la précédente étape), pendant laquelle

nous calculons divers critères statistiques pour détecter un nombre de classes intéressant et construire de bons noyaux initiaux (cf: A.h).

Etape 3 phase d'optimisation de la partition : on effectue de nouveau les Nuées Dynamiques avec ces noyaux "optimaux" pour améliorer la partition finale. Quand la convergence est atteinte (ie quand l'inertie intra-classes cesse de décroître sensiblement) cette partition est alors analysée (description des variables et des classes).

Pour cette troisième phase, il est également possible - l'option existe dans notre programme - d'utiliser à la place des Nuées Dynamiques un algorithme d'optimisation de partitions comme les algorithmes d'Echanges ou de Transferts (FRIEDMAN et RUBIN); on part alors de la partition obtenue avec la coupure de l'arbre hiérarchique de la phase de recherche du nombre de classes. Cette partition étant généralement déjà de bonne qualité ces algorithmes sont alors très efficaces.

Schema de l'algorithme de classification



A.d : La sélection des noyaux initiaux

Pour les techniques d'agrégation autour de centres mobiles, les partitions dépendent fortement des premiers centres choisis et également de l'ordre des éléments à classer. En fait le choix des noyaux initiaux conditionne fortement les résultats obtenus.

Pour sélectionner les noyaux initiaux, nous proposons la démarche suivante:

Le centre de gravité du nuage est choisi arbitrairement comme premier noyau; et on choisi les autres noyaux i en cherchant à résoudre le problème d'optimisation suivant :

$$\text{Max } [\Sigma^{k}_{i=2}\Sigma^{i}_{l=1}\, d(i,l)\,[\,/\,\{Nb\_max - k + 1\}*\{(k.(k-1))\}$$

$$k \leq Nb\_max\_de\_classes\_imposé \quad (30)$$

$$\forall\, i \quad 1 < i \leq n \qquad d(i,G) \leq dmax$$

$$\forall\, i \quad 1 < i \leq n \qquad d(i,G) \geq dmin$$

$$\forall\, i,j \quad i \neq j \qquad d(i,j) \leq dmax$$

$$\forall\, i,j \quad i \neq j \qquad d(i,j) \geq dmin$$

où G désigne le centre de gravité du nuage.

Géométriquement cette stratégie de sélection des noyaux de départ revient à conserver comme noyaux de départ, des points situés entre deux hypersphères dont les rayons sont dmin et dmax; en leurs imposant de surcroît une contrainte de discontiguïté.

Cependant ce mode de sélection impose le choix à priori des rayons de sélection. Pour ce faire on tire au hasard un échantillon d'au plus deux cents éléments de l'ensemble des objets à classer que l'on trie, dmin est alors choisi comme premier quartile, dmax comme troisième.

Enfin, il est certain que cette stratégie de sélection des noyaux pourrait être fort coûteuse en temps cpu avec de grands ensembles de données; ce qui serait évidemment désastreux. Cette étape n'étant qu'un prélimininaire, on a choisi de limiter la résolution de ce problème à un échantillon -tiré au hasard- d'au plus mille objets à classer.

A.e : Les distances utilisées

Pour la classification nous avons utilisé diverses distances, toutes ont fournie une partition intéressante d'un point de vue essentiellement pratique.

Dans toutes les classifications les données sont centrées ; on constate alors les faits suivants:

- avec la métrique identité : $I_{(p,p)}$, les distances euclidiennes favorisent les situations de surcoûts de production très élevés (Fioul et Défaillance).

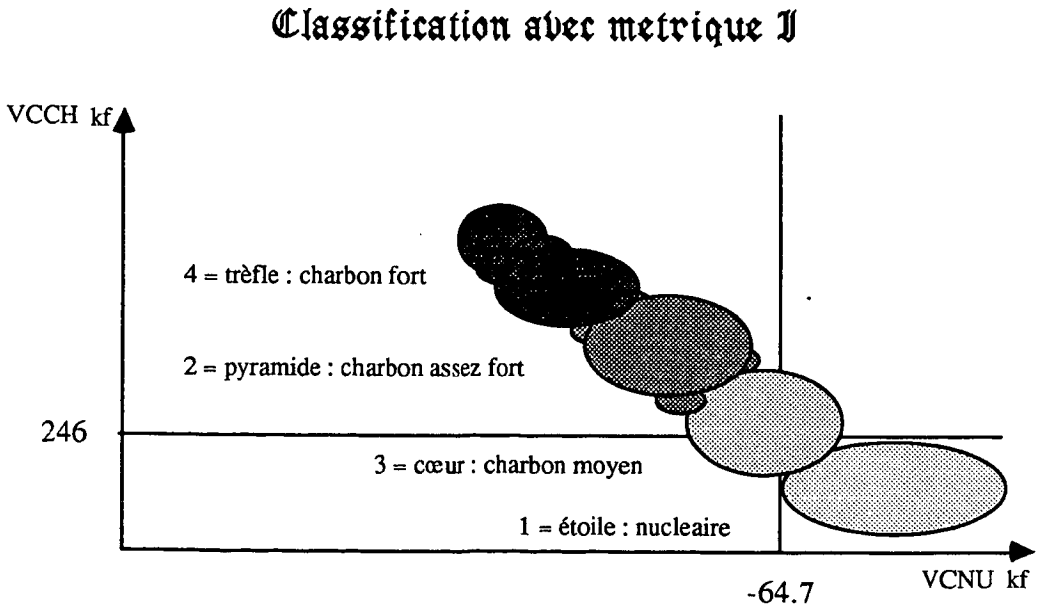
- avec la métrique $1/\sigma^2\, I_{(p,p)}$, les variables deviennent aussi discriminantes les unes que les autres; ce

qui permet de mettre en évidence des classes de situations peu fréquentes notamment celle où l'on constate un turbinage très important avec une utilisation importante du charbon et du fioul (2% des situations)

La métrique Identité ne permettant pas de tenir compte des corrélations qui existent entre nos variables, car dans MEXICO elles vérifient la contrainte suivante: ΔPnucl+ΔPchar+ΔPhydr+ΔPfioul+ΔPdef=0 (l'hypothèse de consommation doit être satisfaite), nous avons choisi d'essayer les métriques de Mahalanobis :

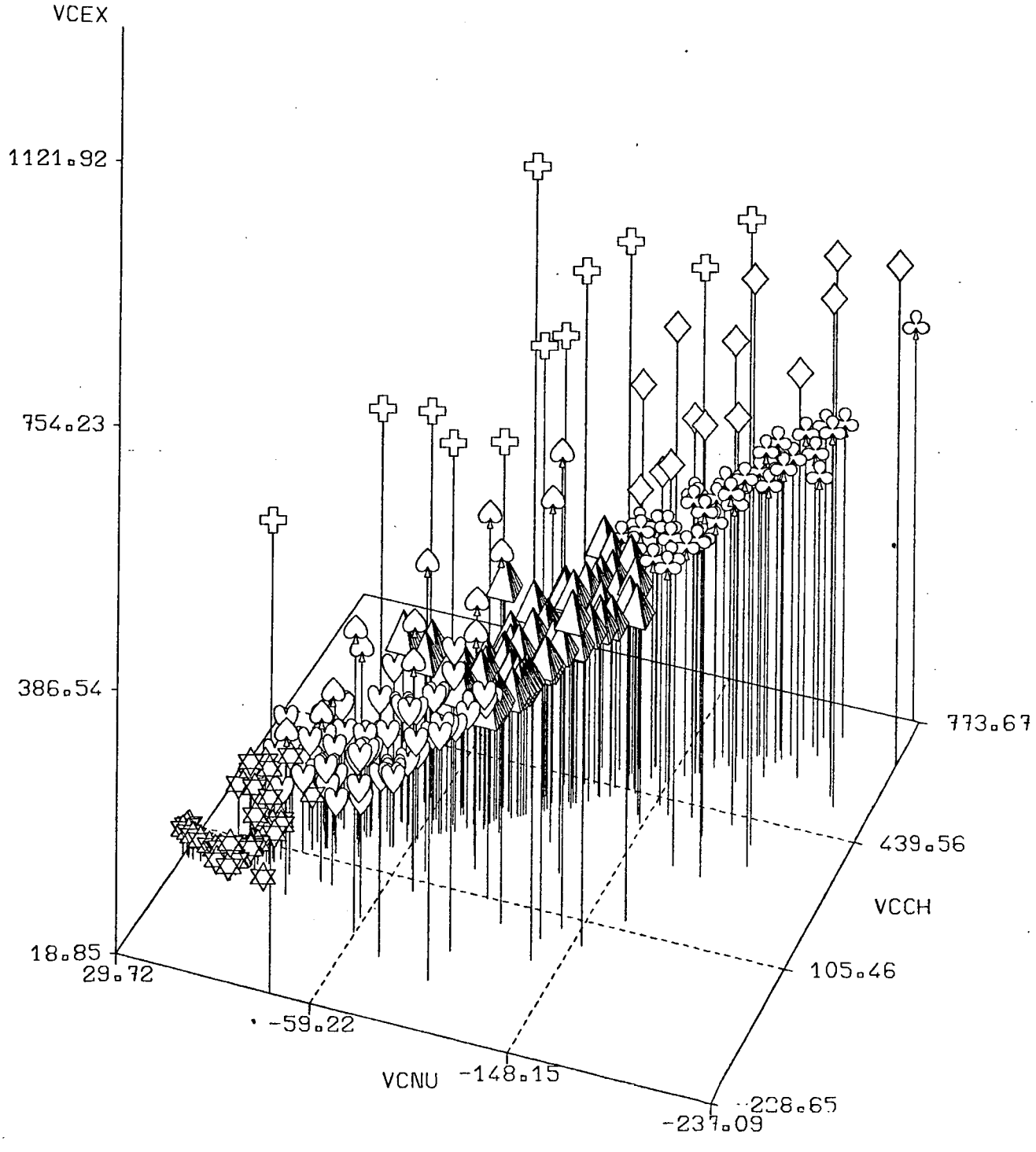
-avec la métrique $V^{-1}_{(p,p)}$ (matrice de covariances totale inverse) les variables sont aussi discriminantes les unes que les autres cependant les variables de surcoût nucléaire et charbon se distinguent quelque peu; on retrouve également les classes de situations peu fréquentes notamment les quelques situations où l'on constate un turbinage très important avec une utilisation importante du charbon et du fioul (comme avec la métrique : $1/\sigma^2 I_{(p,p)}$); mais elle permet également de segmenter les classes en faisant apparaître comme nous le souhaitions des profils: les classes deviennent plus allongées.

-avec la métrique $W^{-1}_{(p,p)}$utilisée exclusivement dans la phase d'optimisation de la partition (matrice de covariances intra-classes inverse) les variables gardent un pouvoir discriminant comparable cependant les variables de surcoût fioul et hydraulique prennent alors plus d'importance contrairement aux autres métriques; mais elle permet également de segmenter les classes en faisant apparaître encore plus nettemment qu'avec la métrique $V^{-1}_{(p,p)}$, des profils. Les classes ne sont plus du tout sphériques comme avec la métrique I mais très nettement "alongées".



Avec la métrique $1/\sigma^2 I$ les classes sont sphériques et représentent les divers paliers du surcôut d'exploitation du thermique total (VCEX) . On retrouve ceci sur le plan VCNU,VCCH de la figure 1.

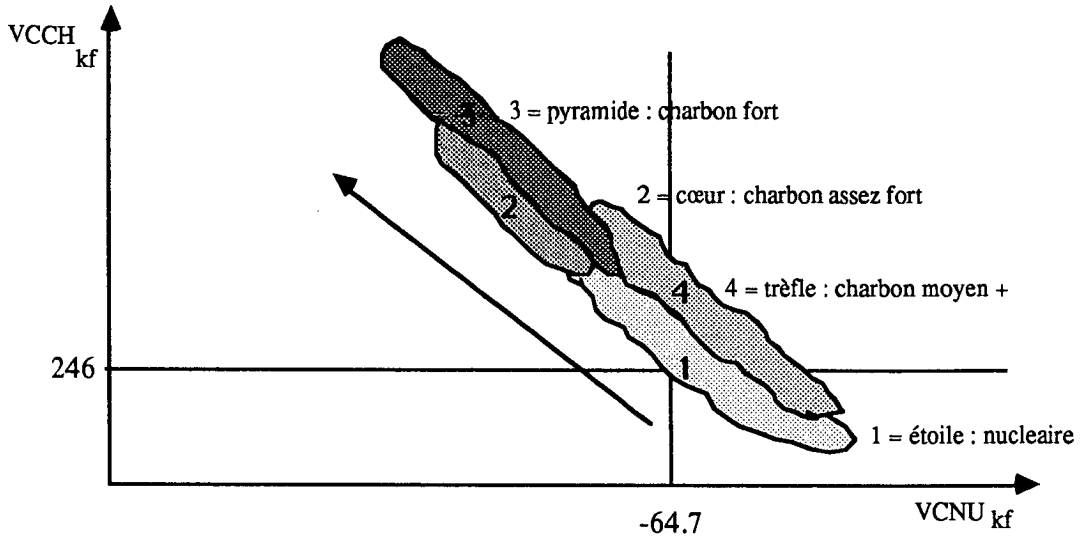
CLASSIFICATION DES SURCOUTS



METRIQUE EUCLIDIENNE I

Figure 1

Classification avec metrique W-1



Avec la métrique W^{-1} les classes sont allongées et soulignent l'existence de problèmes probablement plus liés aux contraintes de réseau (VCEX) . On retrouve ceci sur le plan VCNU,VCCH de la figure 2.

A.f : Les critères de sélection du nombre de classes

Trois critères de jugement de la partition produite à un niveau donné de l'arbre des classifications peuvent être considérés et mis en œuvre.

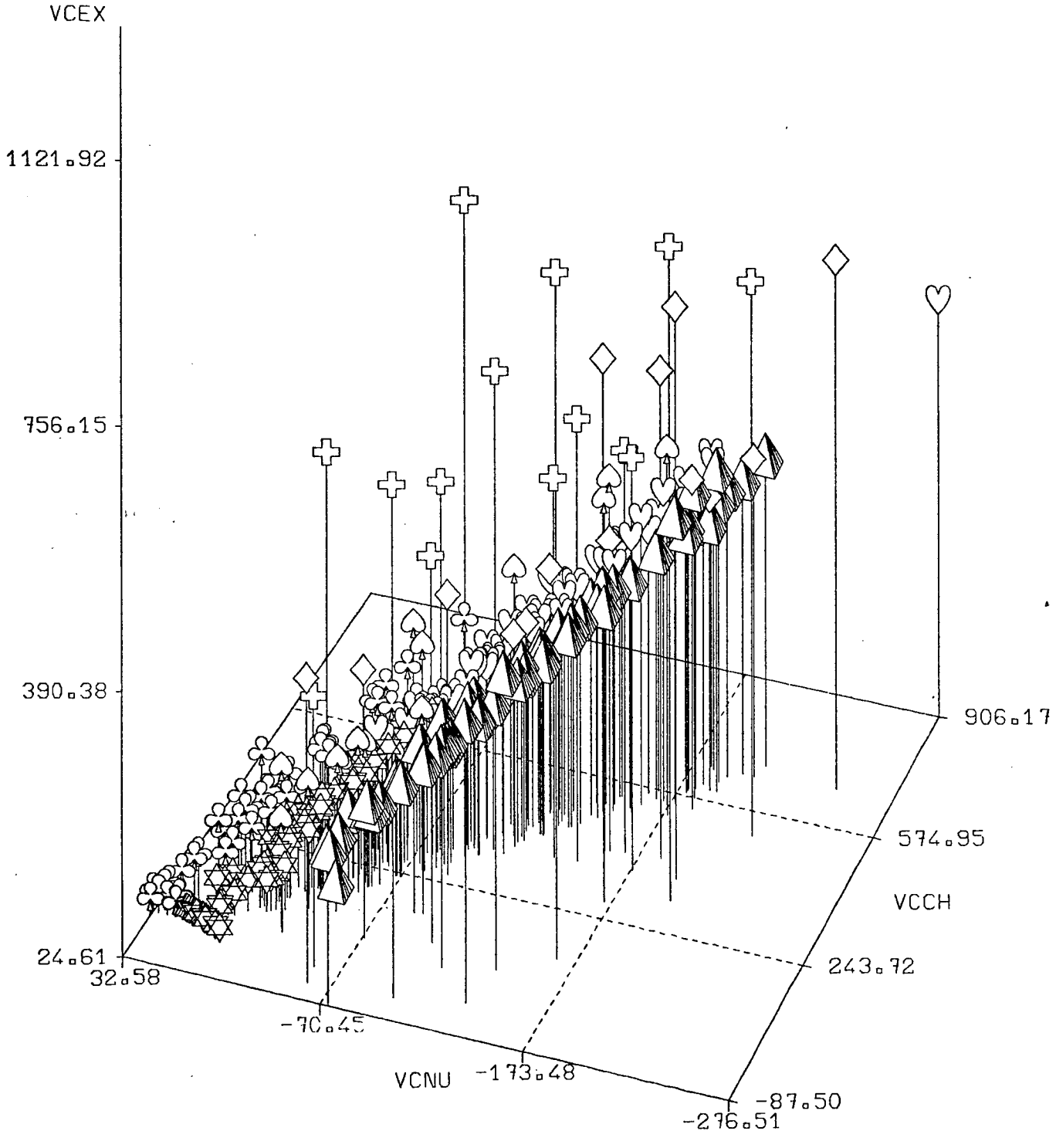
Ces critères sont les suivants. Il s'agit :

- de **l'inertie expliquée** qui cherche à estimer le degré de précision de la partition correspondante et qui est le critère à optimiser dans la plus-part des méthodes de classification. Rappelons en succintement la formule: $\Sigma^k \mu_k \, d\,(G_k\,,\,G)^2$

- d'un **critère cubique de classification** dont la philosophie basée sur un test d'hypothèse cherche à estimer l'existence de classes significatives dans les données. On peut toujours, en effet exécuter un programme de partitionnement automatique non hiérarchique sur des données non classifiables (par exemple, tirées à partir d'une loi de probabilité uniforme), on obtiendra quand même une partition, bien qu'elle n'ait aucun sens. Le critère cherche précisemment à éviter ce genre de problèmes.

Ce critère empirique compare le R^2 de la classification à chaque niveaux de l'arbre hiérarchique à un R^2 théorique calculé avec une formule élaborée empiriquement à partir de données générées par des simulations de Monte Carlo.

CLASSIFICATION DES SURCOUTS



METRIQUE DE MAHALANOBIS B-1

Figure 2

- d'une **statistique locale** (LERMAN 1981) qui cherche à mesurer le degré de dissemblance des éléments que l'on agrège à chaque étape.

Ce critère très général est celui fondé sur la "préordonnance". Si nous désignons par K l'ensemble à classifier qui peut correspondre, soit à l'ensemble V des variables, soit à l'ensemble 0 des objets, la préordonnance sur K est un préordre total sur l'ensemble $L=P_2(K)$ des paires d'éléments de K. Pour ce préordre que nous supposons ici -pour simplifier- un ordre total et strict, le rang d'une paire est une fonction croissante de la ressemblance entre ses composantes, mesurée par l'indice Q de proximité choisi :

$\forall$ (p,q) $\in$ LxL, p < q <=> Q(p) < Q(q)

Nous représenterons dans LxL la préordonnance $\omega(K)$ par son graphe:

gr (ω) = {(p,q) / (p,q) $\in$ LxL, p<q et non q<p pour ω}

Une même partition Π qui jouera ensuite le rôle de P_l ($0 \leq l \leq m$) sera représentée dans LxL par le "rectangle" $R(\pi)$ x $S(\pi)$ ou $R(\pi)$ (resp. $S(\pi)$) est l'ensemble des paires réunies (resp. séparées) par la partition Π. $R(\pi) < S(\pi)$ pour l'ordre quotient.

L'indice brut entre la préordonnance $\omega(K)$ et la partition Π est alors :

$s(\omega , \pi) = \text{card} [\text{gr} (\omega) \cap (R(\pi) \text{ x } s(\pi))]$

Nous opérons une normalisation de cet indice en associant à la partition Π, une partition aléatoire Π^* dans l'ensemble -muni d'une probabilité uniformément répartie - P(n,t) de toutes les partitions de même type cardinal que Π.

La forme la plus simple de l'indice normalisé qui prend le nom de statistique locale est la suivante:

$$S = [s(\omega , \pi) - [r(\pi).s(\pi) / 2] / [r(\pi).s(\pi).[r(\pi)+s(\pi) + 1] / 12]^{0.5}$$

où $r(\pi)$ est le cardinal de $R(\pi)$ et $s(\pi)$ celui de $S(\pi)$.

La suite des valeurs de ce critère local permet de reconnaître quels sont les principaux états d'équilibre dans la synthèse automatique, fournie niveau après niveau dans l'arbre détaillé des classifications emboîtées.

B : LE MODELE LINEAIRE

Avec la classification que nous avons employée pour l'analyse macroscopique des aléas du modèle MEXICO, nous avons utilisé une des nombreuses méthodes d'analyse des données multidimensionnelles. Cette méthode est très intéressante, mais demeure cependant descriptive et non explicative.

Dans le cadre de cette étude, il fallait aller "plus loin" et chercher dans la topologie des liaisons électriques et dans les disponibilités des divers moyens de production, les causes des différents problèmes que la classification avait permis d'identifier. Pour ce faire, nous avons utilisé des méthodes plus explicatives que descriptives, que l'on regroupe sous le nom de : Modèle Linéaire.

B.a : Pourquoi le Modèle Linéaire dans notre application ?

Avec le Modèle Linéaire classique, nous avons essayé d'**expliquer les surcoûts d'exploitation** thermiques et hydraulique en fonction des variables d'entrée du modèle MEXICO : les disponibilités des liaisons électriques et des groupes de production.

Nous avons également essayé d'expliquer les surcoûts d'exploitation en fonction des liaisons en contrainte (une liaison est en contrainte si la puissance qu'elle fait transiter est égale à sa capacité de transport, au delà la liaison risque de fondre), des groupes thermiques chers démarrés et des groupes peu chers ralentis à cause des contraintes sur les liaisons.

Nous avons enfin essayé d'**expliquer les gains marginaux** des liaisons en contrainte en fonction des disponibilités des liaisons et des groupes thermiques.

B.b : l'analyse de l'effet réseau

Cette étape fut abordée par le Modèle Linéaire classique des Moindres-Carrés; nous cherchons dans la topologie des liaisons électriques et dans la disponibilité des groupes de production, l'explication de certaines variables en particulier les surcoûts d'exploitation précédemment étudiés, mais aussi celle des gains marginaux sur les liaisons en contrainte.

L'idée générale est d'écrire le modèle symbolique suivant, les surcoûts d'exploitation fonctions linéaires des disponibilités des liaisons du réseau et des groupes thermiques et des interactions entre ces éléments (interactions précisemment liées au fait que nous traitons un problème d'électrotechnique complexe régit par les deux lois de Kirchoff) ; ce que l'on note ainsi :

$$(VCNU,VCCH,VCHY,VCFG) = \mu + L + S + L * S + \varepsilon$$

avec μ : termes constants des modèles

L : matrice des capacités des liaisons de dimension (1500,238)

S : matrice des disponibilités des groupes thermiques de dimension (1500,148)

L * S : matrice de toutes les interactions entre les liaisons et les groupes

ε : termes résiduels des modèles

(VCNU,VCCH,VCHY,VCFG: surcoûts d'exploitation Nucléaire, Charbon, Hydraulique,Fioul)

L'ajustement d'un tel modèle n'est absolument pas réalisable, même avec un hyper ordinateur, la taille du problème est en effet colossale. Pour s'en convaincre faisons le bilan des éléments explicatifs du modèle.

Pour les capacités des liaisons, nous avons déjà 238 facteurs qualitatifs, pour les groupes 148 covariables , soit déja 386 variables explicatives. Pour les interactions à deux facteurs au sens large, chaque liaison engendre 148 interactions soit un total de 35224. Il faudrait donc pour ajuster statistiquement ce modèle disposer au départ d'un jeu de données de plus de 35600 situations, ce qui n'est pas notre cas; bien que ceci ne soit pas irréalisable. L'impossibilité apparait dans les calculs du Modèle Linéaire : il faudrait alors inverser une matrice de covariances de dimension (35600,35600) , ce qui statistiquement n'aurait absolument aucun sens.

Même en supprimant au départ ,toutes les interactions, il reste quand-même environ 400 paramètres à estimer, nombre qui demeure encore très élévé. Ce n'est plus un modèle irréalisable statistiquement mais cela pose, l'expérience le montre, de sérieux problèmes numériques même sur un CRAY en double précision.

Sachant que tout n'est pas mauvais dans le réseau (heureusement), il convient de chercher au préalable quels sont les groupes de production et les liaisons électriques qui ont une influence statistiquement mesurable sur les variables à expliquer (surcoûts d'exploitation ou gains marginaux). Ensuite on pourra s'il y a lieu, s'intéresser aux éventuelles interactions entre les éléments sélectionnés (interactions entre les liaisons, entre les groupes, entre les liaisons et les groupes).

B.c : La méthodologie d'analyse du réseau

B.c.1 : La sélection des paramètres explicatifs

Pour sélectionner les éléments significatifs du réseau, nous avons utilisé la partition obtenue avec la métrique $1/\sigma^2 I_{(p,p)}$ (voir analyse macroscopique A.a).

Pour sélectionner les liaisons électriques, on peut se poser les questions suivantes:

"pour une liaison électrique, les pannes se sont elles produites avec des fréquences identiques dans toutes les classes, ou au contraire des différences mesurables sont elles apparues ? "

"pour une liaison électrique, les pannes ont elles eu des influences sur les surcoûts d'exploitation, identiques dans toutes les classes ou au contraire des différences mesurables sont elles apparues ? "

Pour la première question, il serait très facile, en théorie, d'y répondre simplement : des *études de fréquences* avec des tableaux de contingence et le test de CHI 2. Seulement nous nous heurtons à un très sérieux problème lié aux données. En effet dans MEXICO les fréquences des pannes des liaisons électriques sont choisis égales à 1%, et dès lors le test de CHI 2 n'a statistiquement plus aucun sens.

Puisqu'il n'est pas possible de répondre à la première question simplement, essayons la deuxième qui est beaucoup plus délicate. Si nous parvenons à répondre à cette question qui revient à estimer l'influence des pannes sur les divers problèmes mis à jour par la classification, nous aurons déjà beaucoup progréssé.

Pour aborder cette question, on pourrait en théorie,utiliser les techniques d'*analyse discriminante* ; cependant avec 238 variables (on travaille alors avec les capacités des 238 liaisons), faire de la discrimination n'a statistiquement aucun sens. Sans compter que l'on a une très faible variation sur chaque variable explicative, qui est liée aux fréquences des pannes.

Finalement nous avons essayé le Modèle Linéaire. Mathématiquement on peut en effet répondre à la question en étudiant le modèle (plus exactement les modèles : il y en a quatre) d'analyse de la variance à deux facteurs et une interaction ci dessous:

$$(VCNU,VCCH,VCHY,VCFG) = \mu + l + classe + l * classe + \varepsilon$$

avec μ : termes constants des modèles

l : vecteur des capacités codées d'une liaison de dimension (1500)

$classe$: vecteur de la partition hors défaillance de dimension (1500)

$l * classe$: interaction entre la liaison et la partition

ε : termes résiduels des modèles

L'interaction $l * classe$ mesure alors pour la liaison l à tester; sur les variables à expliquer (ici les surcoûts d'exploitation), l'existence de différences significatives dans les classes qui sont spécifiques chacune d'un problème particulier.

Pour sélectionner les disponibilités des groupes thermiques de production , on peut se poser la question suivante qui nota bene, n'est pas la seule ni l'unique; mais probablement la plus délicate à traiter :

"pour un groupe de production fixé, pour des niveaux de puissance disponible comparables, les pannes ont elles eu des influences sur les surcoûts d'exploitation, identiques dans toutes les classes ou au contraire des différences mesurables sont elles apparues ? "

Mathématiquement on peut en effet répondre à la question en étudiant le modèle d'analyse de la covariance à deux facteurs et une interaction suivant :

$$(VCNU,VCCH,VCHY,VCFG) = \mu + g + classe + g * classe + \varepsilon$$

avec μ : termes constants des modèles

g : vecteur des disponibilités d'un sommet de dimension (1500)

classe : vecteur de la partition hors défaillance de dimension (1500)

g * classe : interaction entre le sommet et la partition

ε : termes résiduels des modèles

remarque : Toutefois on ne travaille pas sur la disponibilité d'un groupe particulier mais sur la somme des disponibilités des groupes de même type (nucléaire, charbon, fioul & gaz) sur un sommet donné. De ce fait la disponibilité en chaque sommet est distribuée suivant une loi multinomiale.

L'interaction g* classe mesure alors pour le groupe g à tester; sur les variables à expliquer (ici les surcoûts d'exploitation globaux), l'existence de différences significatives de pentes (on sait par expérience que le surcoût d'exploitation global est fortement lié à la disponibilité globale, c'est d'ailleurs principalement pour cela que la sélection des groupes suit une stratégie différente de celle des liaisons),dans les classes qui sont spécifiques chacune d'un problème particulier.

On sélectionne également les liaisons en contrainte, les groupes chers démarrés ou les groupes "pas chers" ralentis, de la même manière que les disponibilités des liaisons (modèles d'analyse de la variance). La sélection des gains marginaux suit quant à elle, la même stratégie que la sélection des disponibilités des groupes thermiques.

Les modèles à deux facteurs et une interaction :

(VCNU,VCCH,VCHY,VCFG) = μ + X + classe + X * classe + ε

utilisés pour la sélection des éléments constitutifs du réseau susceptibles d'avoir une influence statistiquement mesurable sur les surcoûts d'exploitation globaux nucléaire, charbon, hydraulique , fioul; reviennent globalement en fait à une sélection de variables discriminantes comme pourrait le faire un programme de sélection classique, par exemple le programme SELDSC de la bibliothèque SICLA . Cependant ces modèles sont dans cette application, beaucoup plus efficaces pour plusieurs raisons :

1: Ces modèles sont très sensibles . En effet la variable de classe provenant de la classification des variables à expliquer (VCNU--VCFG) , l'effet de ce facteur qualitatif est bien évidemment très fort surtout sur les variables les plus discriminantes (VCNU,VCCH). Le R2 (coefficient de détermination du modèle), rapport de la somme des carrés des écarts expliquée sur la totale est dans ces modèles généralement très proche de 1; en particulier pour les surcoûts nucléaire et charbon où ils sont généralement voisins de 90%. De ce fait la somme des carrés des écarts résiduelle avec laquelle on effectue les tests Fishèriens nécessaires à l'analyse de variance est statistiquement *petite* ; ce qui permet à ces modèles d'être sensibles même à de faibles effets du paramètre X.

Ceci est fort intéressant, en particulier pour la délicate sélection des liaisons électriques dont les taux de pannes sont très faibles. En effet le critère de décision est calculé dans ces modèles d'analyse de variance ou de covariance, sur les variables à expliquer (surcoûts globaux) dont les variances sont importantes comme le

montrent les classifications; alors que dans tous les programmes classiques de sélection de variables discriminantes, le critère de décision (lambda de WILKS par exemple) est calculé avec les variables explicatives.

2: Ces modèles sont homogènes . En effet ils comportent tous le même nombre de variables explicatives : ce sont tous des modèles saturés à deux facteurs et une interaction; ou à un facteur, une covariable et une interaction. De ce fait, pour tous les paramètres X à tester, on peut simultannément estimer l'existence d'un effet global et d'un effet par classe, ce qui est doublement intéressant; à postèriori on peut (si on fait preuve de beaucoup de patience) les comparer entre eux.

3: Ces modèles sont rapides . En effet le programme développé pour cette application traite "en parallèle" les variables à expliquer, ce qui permet une éxécution rapide. Le traitement d'un paramètre prend 1 sec de CPU sur CRAY / XMP. C'est alors le nombre d'éléments à traiter qui rend cette étape délicate (238 liaisons 148 groupes).

B.c.2 : La synthèse des éléments sélectionnés

Après la phase de sélection automatique des paramètres (du réseau) explicatifs, on synthétise l'information extraite par tous les modèles de pseudo discrimination (modèles avec la variable de classe) ; par l'analyse des modèles globaux suivants :

$$(VCNU,VCCH,VCHY,VCFG) = \mu + P + \varepsilon$$

avec μ : termes constants des modèles

P : matrice des éléments du réseau sélectionnés

ε : termes résiduels des modèles

Ce sont les résultats de ces modèles globaux (que l'on étudie automatiquement après la phase de sélection) que nous recommandons de retenir.

Certes, les résultats des modèles de sélection sont évidemment intéressants, mais leur nombre est tel qu'il serait fort long de les étudier tous les uns après les autres. C'est pourquoi nous préferons utiliser les résultats des modèles globaux; modèles qui peuvent être de diverses natures selon les variables introduites pendant la phase de sélection.

Conclusion

Voici donc résumé en quelques pages le travail d'une année pendant laquelle nous avons utilisé les techniques d'analyses des données pour une application industrielle de grande envergure (analyse des résultats du modèle de planification du réseau électrique à très haute tension). Notre recherche ne se veut pas théorique mais plutôt méthodologique; chronologiquement elle se déroule ainsi:

1/ définition d'un objectif de travail, choix des données, utilisation des logiciels disponibles (SAS, MODULAD, SICLA).

2/ Si les résultats sont intéressants, analyse des problèmes rencontrés et définition d'un cahier des charges relatif à la méthode.

3/ Adaptation de cette méthode à notre application.

Bibliographie

CEMBRZYNSKI (1986)
Rapport d'analyse des résultats de MEXICO
Rapport Interne EDF . DER . MOS

DIDAY et Collaborateurs (1979)
Optimisation en Classification Automatique (2 vol)
INRIA

H.P. FRIEDMAN et J. RUBIN
On some invariant criteria for grouping data
JASA 1967, vol 62

IC.LERMAN (1982)
Classification et Analyse Ordinale des Données
Dunod

SR.SEARLE (1971)
Linear Models
John Wiley & Sons

R.TOMASSONE,E.LESQUOY,C.MILLIER (1983)
La régression nouveaux regards sur une ancienne méthode statistique
Masson

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique